Classifying objects in LWIR imagery via CNNs

Rodger, I., Connor, B., & Robertson, N. M. (2016). Classifying objects in LWIR imagery via CNNs. In *Proc. SPIE: Electro-Optical and Infrared Systems: Technology and Applications XIII* (Vol. 9987, pp. 99870-99884) https://doi.org/10.1117/12.2241858

Published in:
Proc. SPIE: Electro-Optical and Infrared Systems: Technology and Applications XIII

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Classifying Objects in LWIR Imagery via CNNs

Iain Rodger,[a,b] Barry Connor [b] and Neil M. Robertson[c]

[a]School of Engineering & Physical Sciences, Heriot Watt University, Riccarton, EH14 4AS, UK;
[b]Thales UK, 1 Linthouse Road, Glasgow, G51 4BZ, UK
[c]School of Electronics, Electrical Engineering & Computer Science, Queen's University, Belfast, BT7 1NN, Northern Ireland

ABSTRACT

The aim of the presented work is to demonstrate enhanced target recognition and improved false alarm rates for a mid to long range detection system, utilising a Long Wave Infrared (LWIR) sensor. By exploiting high quality thermal image data and recent techniques in machine learning, the system can provide automatic target recognition capabilities. A Convolutional Neural Network (CNN) is trained and the classifier achieves an overall accuracy of $> 95\%$ for 6 object classes related to land defence. While the highly accurate CNN struggles to recognise long range target classes, due to low signal quality, robust target discrimination is achieved for challenging candidates. The overall performance of the methodology presented is assessed using human ground truth information, generating classifier evaluation metrics for thermal image sequences.

Keywords: Long Wave Thermal-Infrared, Object Classification, Convolutional Neural Network, Machine Learning

1. INTRODUCTION

Intelligent signal processing capabilities are highly desirable in surveillance tasks for the security and defence sector. Automatic Target Detection (ATD) and Automatic Target Recognition (ATR) are two critical aspects of surveillance based applications. We first address the ATR problem in the LWIR domain by utilising state-of-the-art machine learning methods, the results of which are then incorporated into a larger Automatic Target Detection & Recognition (ATDR) system. This scheme utilises sensor information gathered from a thermal-band sensor.

Motivation: Employing infrared sensor platforms for security related applications is a widespread practice,[1–5] where a thermal sensor is sensitive to the infrared portion of the Electromagnetic (EM) spectrum. The LWIR band existing over the range of $8 - 12\mu$m is emission dominant and ideal for passively imaging hot objects. The utilisation of this spectral band offers increased knowledge of the surrounding environment as thermal imagers effectively *see at night*, providing persistent surveillance capabilities. A typical scenario where such a capability is desirable could be realised as a land reconnaissance vehicle, equipped with an infrared sensor platform and a crew tasked to provide relevant intelligence for a target scene. Ultimately the goal is to improve the overall *situational awareness* by effectively exploiting information gathered from sensors. However, additional information sources increase the burden on a user/operator to quickly process incoming image data and accurately report findings. Extra loads placed on an operator in a stressful, potentially hazardous, environment could have disastrous consequences if a crucial detail is overlooked.

Effective ATDR methods become invaluable as they address this problem scenario by automating the signal processing and alleviate the bulk of the task from a human user. The automatic system could, for example, remove extraneous details leaving only salient regions of interest, or highlight the most important aspects in the surrounding scene prioritised by threat level.[6] Both examples illustrate the system presenting an operator with a vastly reduced information load, but with a greater perception of surroundings, requiring significant effectual

Further author information: E-mail: Iain.Rodger2@uk.thalesgroup.com

automatic signal processing methods. There is an existing body of prior work focused on creating such techniques[1,7,8] which our work indirectly improves upon and advances the field. We achieve this by capitalising on recent machine learning methods to create an object classifier for high quality thermal image data. Thus, we explore prior knowledge in two key areas: the surge in machine learning using Convolutional Neural Networks (CNN) and existing infrared (IR) target classification schemes.

Related Work: Despite the many successes of CNN based applications for colour band imagery, a significant gap remains for thermal band data which this paper addresses. Given the 24-hour sensing capability of a thermal imager (TI) it is ideal for security and defence tasks, where advanced recognition algorithms via CNNs would be highly desirable. Prior work in this area is surprisingly scarce, with only a handful of methods[9–12] utilising CNNs for IR based imagery. When we consider the dearth of publicly available, large scale datasets containing IR imagery, the sparsity of such methods may be explained by the fact CNNs are still recent developments in computer vision, as there are many more classification schemes for thermal band data using older machine learning tools and practices.[1,13,14]

While these works do show impressive performance for IR image classification tasks over a small number of objects, they will be outperformed by deep learning methodologies that enable more descriptive features to be learned.[12] The automatic discovery of better features leads to increasingly effective object recognition performance. Yet even these powerful new techniques are limited by the overall image quality being classified. In our case we are attempting to detect and classify very small targets over long ranges, where descriptive features become limited in effectiveness at differentiating object classes. This is due to the decreasing amount of available information in target image acquisition over long distances, as objects may only be represented as a small number of pixels leading to significant degradation in quality once up-sampled. We shall explore the task of effective ATR using infrared sensor information for such scenarios.

Contributions: Overall, effective ATDR performance in challenging surveillance scenarios is still a highly sought after capability. The use of thermal imaging sensors is increasing which introduces a trade-off between processing complexity and improved scene perception. Our key contributions can be identified as:

1. Creation of a high performance LWIR object classifier using deep learning with CNNs.
2. Deployment of an end-to-end system for effective ATDR tasks. Performance is reported via extensive evaluation using ground truthed sequences.

2. TRAINING A DEEP NEURAL NETWORK

In this section we outline steps taken to create an object classifier for LWIR imagery, using state-of-the-art machine learning methods. Specifically, a deep feed-forward CNN approach is employed. A CNN structure is composed of stacked convolutional layers, hence the term *deep*, followed by one or more fully-connected layers. The input to one layer is determined by the output of its preceding layer. Convolutional stages are essential building blocks in these networks, as they transform input volumes into output activation volumes via convolved filter responses. The spatial structure or location of the image is preserved in activation map computation. Fully connected layers have connections going from all neurons in the previous layer, to each of the neurons in the current layer. Through this linear combination of inputs (i.e. outputs from preceding layer) powerful abstract reasoning is obtained at the expense of spatial information for the image.

2.1 Generating A Training Set

Given the lack of available LWIR training image datasets containing labeled instances of real targets and false alarms, we had to undertake the task of collecting and ground-truthing an image corpus before classifier training was possible. This required the acquisition of large amounts of video data using a high performance TI, choosing

(a) (b) (c) (d)

Figure 1: The catherine MP LWIR variant is shown in isolation (a) and on a multi-modal sensor platform for performing data acquisition tasks (b). A colour image (c) and corresponding LWIR image (d) illustrate a representative rural scene for long range ATDR tasks.

object classes of interest to land defence and defining a preprocessing stage for object instances.

The Catherine MP LWIR is a state-of-the-art TI produced by Thales[15] and we use it to collect data for training-set generation. The Catherine MP LWIR uses an integrated detector cooler assembly which comprises a 640×512, 20μm pitch quantum well infrared photodetector (QWIP) array, sensitive to long wave infrared radiation at wavelengths of 8μm to 12μm at a frame rate of 100 Hz. This imager is presented in Figure 1 and can be seen deployed in a multi-sensor platform. Note that all experiments presented in this paper only concern signal processing applications for the thermal band, but colour imagery will be shown in some examples to provide clarity. The imagery collected using this high calibre TI will be of sufficient quality to allow the successful application of deep learning methods for thermal band data. We train over a select number of object classes.

2.1.1 Object Class Designation

The motivation for this work and subsequent experiments is driven from a surveillance perspective in land defence scenarios. Thus, the object classes we are interested in are commonly found in urban and rural scenes. We choose six classes to examine: people, land-vehicle, helicopter, aeroplane, Unmanned Aerial Vehicle (UAV) and false alarm (FA). Example instances for each class are presented in Figure 2 demonstrating the kind of images comprising the training and test set. It should be noted that there are five distinct objects and one null hypothesis class, the false alarm. Including this null class is crucial as it allows incorrect target candidates from the ATD process to be successfully rejected. The images that compose the false alarm class are much harder to quantify than the five distinct objects, as it can theoretically include anything outside the small set of objects. We include examples of things in scenes that typically generate false alarms for detectors in LWIR sequences, such as edges of moving clouds, bushes, branches, corners of building etc.

2.1.2 Data Preprocessing

Having defined what object classes are of interest, the captured thermal sequences could then be examined to create a training and test set. The Catherine MP outputs 14bit video data and object crops are taken from each sequence. The image crops resolution varies for each example due to the nature of each objects physical appearance. The CNN framework we employ must have input images with exactly the same dimensions and data is required to be single-precision floating-point format, i.e. pixels are in the range [0 1]. After obtaining enough training examples for each object class, as well as a separate test set, all of the data is transformed to single-precision with each image crop re-sized to an array of size $m \times m$, where $m = 256$. The process of adapting the spatial resolution for each object crop manifests itself by slightly skewing object crops that have more rectangular shape. An example of this is illustrated in Figure 2, on the bottom row of the pedestrian class.

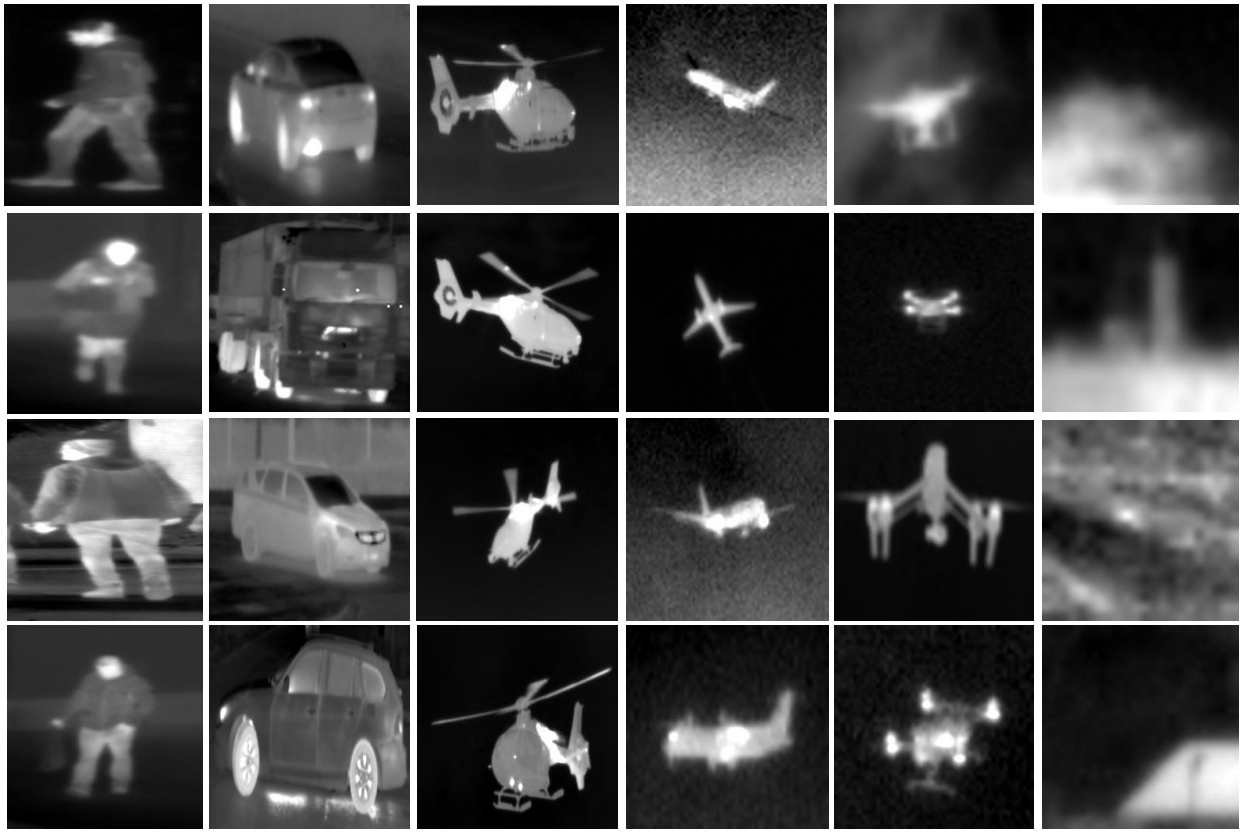

Figure 2: Training examples for each object class, cropped from Catherine MP LWIR imagery. Instances shown in column (a) highlight people from various poses. Land vehicles are observed in column (b), showing not only different pose/viewpoints but also intra-class variation. Instances of helicopters can be found in column (c), with aeroplanes present in (d). Lastly, column (e) illustrates UAV examples while column (f) highlights various false alarm instances.

Lastly, we also median filter each crop using a 3×3 kernel. After successful data preprocessing the classifier can be trained and evaluated.

2.2 Training Phase

The constructed dataset contains ≈ 11000 LWIR object instances sampled over the 6 object classes, containing 5 real targets and a false alarm class. The set is split in a 90:10 ratio to form the training and test set, where the test set will be effectively unseen during the training process and not affect the weight learning in any way. We define our network architecture and perform the training phase using the python based deep learning frameworks, Theano[16] and Lasagne.[17]

The general approach to training a deep network for extensive datasets is to adopt a network architecture similar to those reported by Krizhevsky,[18] as they are demonstrably effective at the task. However, these network structures are for tackling truly vast datasets with training instances numbering in the hundreds of thousands to millions, with potentially thousands of target output classes. We are only dealing with a training set of ≈ 10000

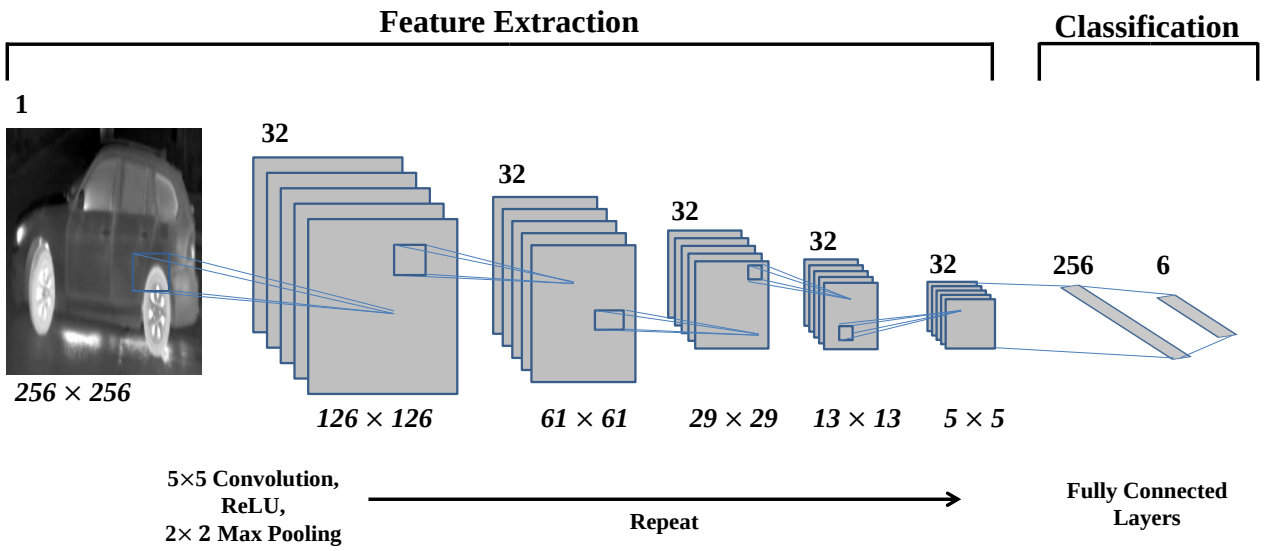

Figure 3: An illustration of our network architecture, showing how data would propagate through the network. Every convolutional layer is composed of 32 filters producing 32 feature maps, shown as image planes here. Each convolutional process is followed by a ReLU and maxpooling operation, which can be thought of as a *block* of processes. This convolutional process block is repeated 5 times from input image until we reach the dense, fully connected layers. The dimensions of each convolutional layer output volume is also provided. For example, the first output volume is composed of 32 feature maps of size 126×126.

examples over 6 output classes. Our network architecture reflects the difference in scale by preserving the overall depth and sequential structure, but removes the width when compared to the Krizhevsky[18] architecture. The network structure we employ is straightforward and highly symmetric. It functions by learning a generisable representation of the data, i.e. rich features, through five convolutional blocks and two fully-connected layers. A diagrammatic overview of the network showing the details of each layer is presented in Figure 3.

The input layer weights are initialised using the Glorot scheme[19] and every convolutional layer is composed of 32 filters. These filters are of size 5×5 and are convolved across the image with a stride length of 1 to generate activation maps. Each convolutional block is followed by applying a non-linear activation function and a pooling step. We use the Rectified Linear Unit[20] function (ReLU) $\phi(x) = max(0, x)$ as the non-linearity, applied to the filter response activation maps, followed by pooling with filter size 2×2 and stride length 2. The pooling operation acts as a downsampling in spatial resolution. The last convolutional stage feeds into a dense, fully-connected layer with 256 units and 50% dropout[21] is applied to layer inputs. Again, the ReLU operation is applied here. The final network output layer is a softmax function,[22] also known as a normalised exponential operation, with 6 units. This final normalising layer has one unit for each object class and outputs a probability distribution over all classes which sum to one. The probabilistic interpretation of the softmax classifier function is:

$$P(y_i|x_i) = \frac{e^{f_{y_i}}}{\sum_j e^{f_j}} \tag{1}$$

which tells us the normalised probability assigned to each label y for each input image x, where f is an output score vector. With the network structure fully defined the data can be passed in for training after setting the optimisation strategy.

Training & Optimisation: There are a few key considerations to examine when training a CNN. The entire process can be summarised as a non-convex optimisation problem, where the goal is to minimise an objective

Classifier Output

Ground Truth \ Class	C1	C2	C3	C4	C5	C6	All
C1	197	2	0	0	0	0	199
C2	5	301	2	0	0	1	309
C3	0	1	105	0	0	0	106
C4	0	0	0	59	0	0	59
C5	0	0	0	0	77	4	81
C6	1	2	0	3	2	238	246
All	203	306	107	62	79	243	1000

6 Class Accuracy = 97.7%

(a)

Classifier Output

Ground Truth \ Class	C1	C2	C3	C4	C5	C6	C7	All
C1	157	1	0	0	1	2	0	161
C2	4	249	0	0	0	4	0	257
C3	0	3	80	0	0	0	0	83
C4	0	0	0	53	0	0	0	53
C5	1	0	0	0	53	10	1	65
C6	3	1	0	5	4	185	3	201
C7	0	0	0	0	0	0	180	180
All	165	254	80	58	58	201	184	1000

7 Class Accuracy = 95.7%

(b)

Figure 4: Test results generated by trained CNN over 1000 unseen LWIR examples for 6 & 7 target classes. The overall CNN accuracy is 97.7% for 6 classes, which drops slightly to 95.7% with the presence of an additional 7th class. Ground truth classes are rows, classifier outputs are columns. The main diagonal reveals classifier performance. **C1** is the person class, **C2** is land vehicle, **C3** is helicopter, **C4** is aeroplane, **C5** is UAV and **C6** is the false alarm class. The additional $7th$ object **C7** is the long range target class .

function. In other words we alter the weights of the network accordingly until the error is sufficiently low. This can be achieved using what are now standard practices. We employ categorical cross-entropy loss as the objective function, which is the common choice for multi-class problems and softmax outputs. Given an appropriate objective function an optimisation strategy is required, where we utilise the standard practice of backpropagation in conjunction with the gradient-descent optimiser Adagrad.[23] Backpropagation allows the calculation of gradients for the cross-entropy loss function, with respect to global network weights. The computed gradient is then used by the gradient-descent algorithm Adagrad to allow network weight updates, where the goal is to minimise the overall error or loss function.

Let us suppose that our objective function is $F(\theta)$, where θ is the model parameters. Straightforward gradient descent minimises $F(\theta)$ by updating parameters in the *opposite* direction to the gradient, given by $\nabla_\theta F(\theta)$. There is also an associated learning rate η responsible for the step size when descending the gradient slope. This parameter is very important to the entire process as there is the risk of overshooting the minimum if the step size is too big, or conversely if it is too small the time to train the network can be very long. By choosing Adagrad we can avoid manually trying to tune the hyperparameter η as learning rate updates are adaptive to the parameters θ, meaning we can set it large enough during the start of the process and still be assured of locating minima. If we observe $d_{t,i} = \nabla_\theta F(\theta_i)$ as the gradient of the objective function for parameter θ_i at time step t, the update rule for each parameter at time t can be given by $\theta_{t+1,i} = \theta_{t,i} - \eta \cdot d_{t,i}$. Finally, the Adagrad update rule is computed as :

$$\theta_{t+1,i} = \theta_{t,i} - \frac{\eta}{\sqrt{D_{t,ii} + \epsilon}} \cdot d_{t,i} \tag{2}$$

which effectively shows us how the algorithm modifies the learning rate η, for θ_i at time t, using previous gradient calculations. The term $D_{t,ii}$ is a diagonal matrix where the main diagonal elements i, i are the sum of squares of the gradients and ϵ is a small smoothing variable to circumvent potential zero divisions. Having created a labeled LWIR object dataset, defined the network architecture and set appropriate optimisation functions, the CNN can be successfully trained. Classifier evaluation results over the unseen test set of 1000 examples is presented in Figure 4 and an example classification demonstration is provided in Figure 5. We can now explore the application of this trained network in an end-to-end system for long range ATDR.

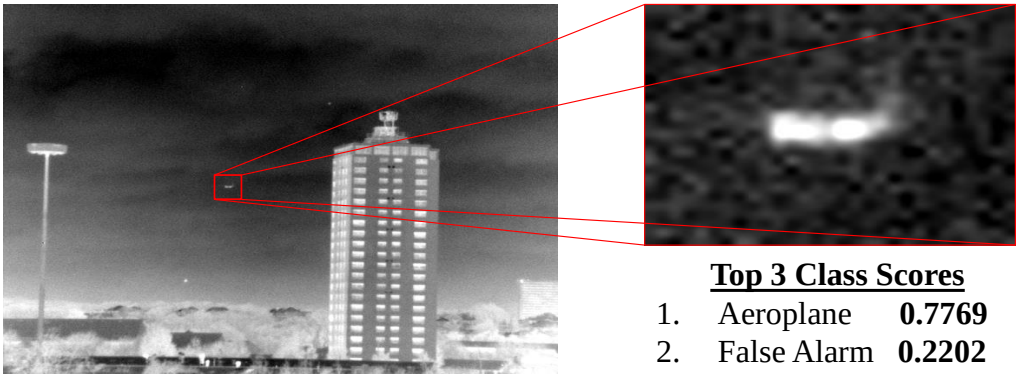

Figure 5: A demonstration of the trained CNNs recognition capabilities towards LWIR imagery. Image (a) provides scene context containing the target, while image (b) is the upsampled target representing what the CNN will classify.

3. LONG RANGE AUTOMATIC TARGET DETECTION & RECOGNITION

An initial end-to-end ATDR system requiring LWIR input data is described and examined in this section. Algorithms are designed around real-world, *unsterilised* data, collected from a static surveillance platform in varying environments. The system design is reflected in the layout of this section and contains three key elements. Given an input data stream the first stage is to generate candidate targets via an ATD algorithm, briefly described in Section 3.1. These candidate detections are then passed to the trained LWIR object classifier for ATR, outlined in Section 3.2. Output probability scores for each detection are then passed into an overarching contextual framework, explained in Section 3.3, utilising prior target knowledge with spatial and temporal location to affect final class scores. The system is initially developed using mid to long range sequences for testing before being fully evaluated over several, longer range surveillance sequences which have been manually ground truthed. By developing and testing over initial examples, we certify the algorithmic structure and identify the need to introduce a new long range target class into the classifier. The additional class enables effective filtering of false alarms from real targets, allowing the scene to be perceived and reported in a meaningful way.

3.1 Autonomous Target Detection System

The central theme of this work is concerned with enhancing overall ATR performance via CNNs, meaning the choice of ATD algorithm is of small significance as long as it can detect targets. Typically with thermal data some form of hotspot detection is employed to generate target regions, for example the method presented by Teutsch.[13] Autonomous target detection is performed only on the thermal image feed from the Catherine MP LWIR. We use a proprietary Thales algorithm for this task, which is capable of localising targets from short to very long ranges. The crucial step is to ensure that any candidate targets are preprocessed in the same manner as described in Section 2.1.2, as target images need to be of the same dimension and format the CNN was trained over. The processed samples for each target are passed forward to the trained CNN for classification, along with localisation information for the temporal context framework.

3.2 Initial Long Range Target Classification

Using Thales' ATD algorithm on thermal sequences we can generate candidate targets to be classified with the trained CNN, giving an early-stage ATDR scheme. The first step for developing an effective end-to-end system is to test this ATDR process on some initial sequences, captured using the multi-sensor set-up shown in Figure

(a) (b) (c)

Figure 6: Short Range Sequence *Country Road*. The top row illustrates registered colour band imagery, showing a person walking toward the sensor platform along a rural path. The bottom row is the corresponding LWIR imagery showing the same scene. ATDR is performed only on the thermal data stream, but bounding boxes are shown on both modalities for clarity.

1. The first test scenario is a short to mid range clip in a rural environment, where a person walks towards the sensor platform as illustrated in Figure 6. The overall evaluation results for this test sequence, *Country road*, are presented in Figure 8, confirming the effectiveness of our CNN for LWIR target classification. This is perhaps unsurprising given the training imagery is of similar quality to the targets from the *Country road* sequence. The second test scenario is a much longer range sequence, *The Braes*, which we again use to test the early-stage ATDR scheme. The imagery from this sequence is illustrated in Figure 7.

As we can see in Figure 7, ATDR for this scenario is very challenging. The target is very small in resolution and low quality. We rather naively apply the initial ATDR scheme to this long range scenario and examine the classification results. Somewhat unsurprisingly the CNN is not capable of assigning correct class labels to targets of such low image quality, highlighted by the less than desirable accuracy and confusion matrix results presented in Figure 8. However, upon closer examination it appears the network can actually differentiate between false alarms and real targets with resounding accuracy, it just gets the object class consistently incorrect. Given the network has not seen images of such low quality in the training process, we capitalise on this capability by gathering long range training examples (similar to subfigure (c) in Figure 7) and retraining the CNN. The network is trained with added long range target imagery using the same parameters and architecture as summarised in Figure 3, except the final fully-connected layer outputs over 7 units instead of 6 to account for the additional class. Output validation results from retraining the LWIR CNN are presented in Figure 4, subfigure (b).

The successful discrimination of real targets and false alarms is achievable via the introduction of a long range target class. However, we still don't know the actual object class of real targets but only that it is an object of interest. The key benefit of introducing a long range object class is that it enables effective target reporting. Unresolvable candidates can now be confidently classified as a long range object of interest and are not incorrectly misclassified as another target class, whilst still being differentiated from false alarms. Recognition of such small targets also implicitly tells us something about the observed scene structure as there must a significant range involved. Ultimately, we exploit infrared information for initial classifications and use spatial/temporal context to enhance the process.

(a) (b) (c)

Figure 7: Long range sequence *The Braes*. Image (a) is a colour band image registered with central LWIR image (b). Both illustrate a candidate detection in the top, central portion of the image. The target is bounded by a red box. Image (c) is the target upsampled from the LWIR image information. Note that the detection algorithm only operates on the thermal image, but is shown on both colour and thermal for clarity.

Classifier Output

Ground Truth	C1	C2	C3	C4	C5	C6	All
C1	2196	77	0	0	0	0	2273
C2	0	0	0	0	0	0	0
C3	0	0	0	0	0	0	0
C4	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0
C6	0	2	0	0	0	525	527
All	2196	79	0	0	0	525	2800

Country Road Total Accuracy = 97.18%

(a)

Classifier Output

Ground Truth	C1	C2	C3	C4	C5	C6	All
C1	0	0	0	0	0	0	0
C2	0	0	0	0	1037	1	1038
C3	0	0	0	0	0	0	0
C4	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0
C6	0	0	0	0	66	1296	1362
All	0	0	0	0	1103	1297	2400

Braes Total Accuracy = 54%

(b)

Figure 8: Initial test results for two sequences. Subfigure (a) is output classification results for the *country road* sequence. The only classes present are **C1** & **C6** which are person and false alarm respectively. Subfigure (b) is output classification results for the *Braes* sequence. The only classes present are **C2** & **C6** which are land vehicle and false alarm respectively.

3.3 Temporal Aggregation

Although CNNs are generalisable and robust at classification tasks, incorrect classifications are still present. One possible method to address this is to informally track targets through a sequence and aggregate the output CNN scores over time, effectively squashing any erroneous probabilities. Considering the aim is to classify targets at long range, we can achieve temporal aggregation without implementing a tracking algorithm as far away targets move very little on the image plane. This simple heuristic complements the long range target class which implicitly implies the observed object exists at a range that is unresolvable. Moreover, traditional tracking algorithms tend to perform data association via kinematics of targets, without considering any aspects of object recognition

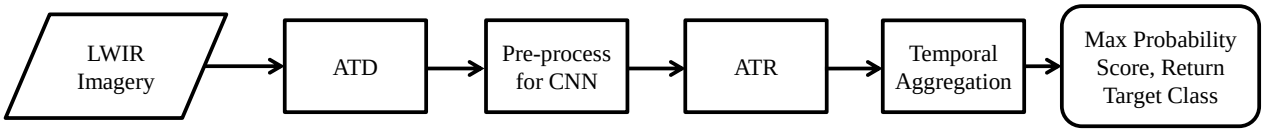

Figure 9: General overview of the ATDR algorithm stages. Given LWIR input data, candidate detections can be generated via an ATD process. These candidates are fed to the trained CNN, where the output score vector is temporally aggregated and reweighted. Maximum probability returns output target class.

which our method incorporates. To implement this we formulate the problem using Bayes theorem to exploit the spatial and temporal relationship of targets in a video sequence. This can be summarised as trying to determine the probability of each object class Obj given previous classifications and detection locations through time T.

We create a circular buffer of detections with corresponding output CNN scores, where each output is a 1-dimensional vector or array of probability class scores $\mathbf{x}_{cnn}$. When a new detection and subsequent CNN score is acquired we can treat it is as an initial confidence of class values for that detection, which is *prior* knowledge $P(Obj)$. Using this we determine the posterior $P(Obj|T)$ by finding the closest spatial match in the circular buffer of detections, which serves as our likelihood function $P(T|Obj)$ to aggregate CNN class scores. To compute the likelihood we search the previous detection locations in the buffer, finding the closest detection on the image plane in terms of Euclidean distance det_{E_d}. If the nearest match is less than or equal to a defined distance threshold $Thresh_{E_d}$, i.e. spatially close, $P(T|Obj)$ becomes the corresponding CNN output vector containing all class scores. If the detection has no match below $Thresh_{E_d}$, current CNN output is unaffected. This condition is summarised as:

$$P(T|Obj) = \begin{cases} \mathbf{x}_{cnn}, & \text{if } det_{E_d} \leq Thresh_{E_d} \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

This gives us Bayes theorem as $P(Obj|T) \propto (T|Obj)P(Obj)$, which effectively describes how to update current target classes based on previous classifications, as well as spatial and temporal observations. The process then moves onto the next acquired detection and corresponding CNN scores, updating the circular buffer as required. By propagating through the detection sequence in this manner, the CNN scores are temporally aggregated and any spurious classifications diminished.

3.4 Final ATDR System

We can piece these components together for an effective end-to-end system and evaluate using real world, long range data. A graphical summary of how the key stages are linked is shown in Figure 9. For an input LWIR data stream we generate candidate targets using Thales's proprietary ATD algorithm. The corresponding image data for each target is upsampled, using the same process steps as employed for generating CNN training data. Target crops can then be fed into the trained CNN and output classes reported. The final process is a temporal aggregation scheme that employs a Bayesian framework and spatial detection information, affecting current target class scores via prior information. The resulting class scores are normalised so all entries sum to 1, the max of which provides the top class result used to evaluate the system.

Classifier Output

Ground Truth / Class	C1	C2	C3	C4	C5	C6	All
C1	2205	68	0	0	0	0	2273
C2	0	0	0	0	0	0	0
C3	0	0	0	0	0	0	0
C4	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0
C6	0	2	0	0	0	525	527
All	2205	70	0	0	0	525	2800

Country Road : CNN + Temporal = 97.5%

(a)

Classifier Output

Ground Truth / Class	C1	C2	C3	C4	C5	C6	All
C1	0	0	0	0	0	0	0
C2	0	0	0	0	1037	1	1038
C3	0	0	0	0	0	0	0
C4	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0
C6	0	0	0	0	35	1327	1362
All	0	0	0	0	1072	1328	2400

Braes: CNN + Temporal = 55.29%

(b)

Classifier Output

Ground Truth / Class	C1	C2	C3	C4	C5	C6	C7	All
C1	0	0	0	0	0	0	0	0
C2	0	0	59	0	223	163	4701	5146
C3	0	14	749	0	5	5	0	773
C4	0	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0	0
C6	0	3	0	0	82	2696	50	2831
C7	0	0	0	0	0	0	0	0
All	0	17	808	0	310	2864	4751	8750

Rural Long Range: CNN Acc = 39.4%

(c)

Classifier Output

Ground Truth / Class	C1	C2	C3	C4	C5	C6	C7	All
C1	0	0	0	0	0	0	0	0
C2	0	0	50	0	222	137	4737	5146
C3	0	13	751	0	2	7	0	773
C4	0	0	0	0	0	0	0	0
C5	0	0	0	0	0	0	0	0
C6	0	3	0	0	84	2701	43	2831
C7	0	0	0	0	0	0	0	0
All	0	16	801	0	308	2845	4780	8750

Rural Long Range: CNN+Temporal Acc = 39.5%

(d)

Figure 10: Confusion matrices and overall accuracy results are presented for classification experiments over three sets of data sequences. **C1**= person, **C2** = land vehicle, **C3** = helicopter, **C4** = aeroplane, **C5** = UAV, **C6** = false alarm and **C7** = long range target class . Matrices (a) + (b) are the updated results using CNN with temporal data association. Matrix (c) shows the simplest case of the CNN applied to ATD candidates for challenging, long range rural test data. Matrix (d) presents the temporal variation evaluation results of the ATDR system applied to the same sequence.

4. LONG RANGE PERFORMANCE EVALUATION

The final ATDR system is comprehensively evaluated using three sets of data sequences. The initial ATDR development sequences, *Country Road* and *The Braes*, are re-evaluated with the complete end-to-end system including the temporal aggregation scheme. Following this, we evaluate unseen challenging long range data sequences collected in a rural location, illustrated in Figure 1. These rural scenes contain two main object classes, land vehicles and helicopter, as well as false alarms to classify. All detections generated via the ATD algorithm are human ground-truthed to provide target classes.

Two combinations of the ATDR system are possible and we evaluate each of them, obtaining overall accuracy results and corresponding confusion matrices shown in Figure 10. The updated evaluation results utilising the temporal scheme for *Country Road* and *The Braes* are included as well as results for the challenging, unseen rural sequences. The simplest variation is *CNN*, where we apply only the trained CNN to targets output from ATD. The other variation *CNN+Temporal* applies the trained CNN and temporal aggregation scheme outlined in Section 3.3. The overall classifier accuracy for each combination is provided by $Acc = \frac{Tr(cfm)}{n_{cfm}}$, where $Tr(cfm)$ is the trace of the confusion matrix and n_{cfm} is the total number of elements in the confusion matrix. We classify a total of 8750 candidate long range targets from the rural scene dataset for the final experiment.

5. DISCUSSION

The validation confusion matrices present in Figure 4 highlight the accurate classifier obtained via our CNN training scheme. Both overall accuracies are greater than 95% , which is remarkable given the addition of a seemingly uninformative $7th$ class for long range targets. The CNN structure of the classifier is highly symmetrical and deep, which appears to work for our training dataset. Although we based our convnet architecture around Krizhevsky's successful ImageNet design, it was a little surprising it translated so well given the network

is designed for much larger scale learning. ImageNet is > 1 Million images over 1000 classes whereas our training set is approximately 10000 images over 7 target classes. The ratio of 1000 images per class is roughly preserved and may be one reason why the network structure works well in both scenarios. The depth and pooling operations also help combat overfitting to the training dataset. In any case, the results obtained for LWIR target classification are remarkable and to our knowledge have never been shown before at this accuracy, or for the range of object classes we test over.

Despite the effectiveness of the trained classifier, initial experiments presented in Section 3.2 highlight the ill-suited nature of CNN-only based classification of objects. Simply throwing a large scale machine learning approach at the long range problem isn't viable for producing reliable target classes. However, these initial experiments identified the capability to differentiate false alarms from real targets, which is extremely useful and we capitalise on it by retraining our CNN with an additional class. By doing so we can confidently report unresolved targets at range, increasing the overall situational awareness. The results of our final ATDR system, guided by the trained CNN and Bayesian temporal aggregation, are presented in Figure 10. The experiments evaluate the end-to-end ATDR system over challenging long range sequences, generating 8750 target classes to examine, as well as a re-examination of the original test sequences first presented in Section 3.2.

It is immediately clear from *CNN* and *CNN+Temporal* total accuracy results that the CNN is robust at providing accurate target class information at short to mid range. However, both variations are incapable of discerning the correct land vehicle classes at longer ranges, as shown in Figure 10 (c) + (d). This will simply be due to the low signal quality at such long ranges. Furthermore, the temporal aggregation function appears to have a negligible positive affect, only raising the overall classifier accuracy by a meager $0.1 - 1\%$ across all cases. This may be explained by the fact that CNN output scores are very rarely *on the fence*. In other words they are very strong, consistently approaching $> 90\%$ even when incorrect. Thus, even with temporal aggregation it makes it very hard to diminish the odd, erroneous classification probabilities. Out of thousands of instances we only manage a handful of correct reclassifications, which is reflected in the negligible performance gain. These gains, while modest, are mostly achieved by better discrimination of false alarms.

The obvious drawback of the system as a whole is the inability to correctly classify long range targets. Despite using state-of-the-art machine learning techniques and a sufficiently large dataset of high quality TI images, it is simply not enough to solve the challenge of long range target recognition. However, if we examine the confusion matrices for the rural scene shown in Figure 10 (c) + (d), the majority of long range targets are clearly differentiated from other classes. This bolsters confidence in target reporting capabilities and if a sufficient mechanism existed to switch the long range class to the correct land vehicle class, the overall accuracy would be significantly higher. This highlights a gap that needs addressed and presents a future research avenue to be explored. Despite this disadvantage, what we do gain is a robust ATDR system for short to mid-range scenarios capable of day and night operation, as well as confident target discrimination for long range scenarios.

6. CONCLUSION

We have presented a complete ATDR system for improved target detection and recognition capabilities in surveillance scenarios using LWIR data. The combination of labeled CNN training dataset instances and ATDR test sequences represent over 20000 human ground truthed examples. The labeled examples are used to generate all outcomes presented in this work. Obtaining such results was achievable by initially adopting state-of-the-art machine learning methods to create a highly accurate LWIR target classifier via CNNs, demonstrating robust recognition across a range of objects in LWIR imagery. The approach is entirely data driven allowing additional information to be incorporated easily, either to classify a new target or improve the current system. We also demonstrated the ability to discriminate between real targets and false alarms to a high degree of accuracy, which can be utilised to improve false alarm rates of an ATD process. Building on this capability, the described ATDR system could potentially be deployed in a reconnaissance scenario and alleviate the burden on human operators via effective target reporting. Overall, the approach should generalise very well to ATDR tasks in the security and defense domain, as well as outside this realm.

REFERENCES

[1] Breckon, T. P., Gaszczak, A., Han, J., Eichner, M. L., and Barnes, S. E., "Multi-modal target detection for autonomous wide area search and surveillance," in [*SPIE Security+ Defence*], 889913–889913, International Society for Optics and Photonics (2013).

[2] Clapés, A., Reyes, M., and Escalera, S., "Multi-modal user identification and object recognition surveillance system," *Pattern Recognition Letters* **34**(7), 799–808 (2013).

[3] Leykin, A. and Hammoud, R., "Robust multi-pedestrian tracking in thermal-visible surveillance videos," in [*2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*], 136–136, IEEE (2006).

[4] Dickson, C. N., Wallace, A. M., Kitchin, M., and Connor, B., "Vehicle detection using multimodal imaging sensors from a moving platform," in [*SPIE Security+ Defence*], 854112–854112, International Society for Optics and Photonics (2012).

[5] Connor, B., Letham, J., Robertson, N., and Carrie, I., "Scene understanding and task optimisation using multimodal imaging sensors and context: a real-time implementation," in [*SPIE Defense, Security, and Sensing*], 80120A–80120A, International Society for Optics and Photonics (2011).

[6] Ratches, J. A., "Review of current aided/automatic target acquisition technology for military target acquisition tasks," *Optical Engineering* **50**(7), 072001–072001 (2011).

[7] Zhang, H., Nasrabadi, N. M., Zhang, Y., and Huang, T. S., "Multi-view automatic target recognition using joint sparse representation," *IEEE Transactions on Aerospace and Electronic Systems* **48**(3), 2481–2497 (2012).

[8] Sun, J., Fan, G., Yu, L., and Wu, X., "Concave-convex local binary features for automatic target recognition in infrared imagery," *EURASIP Journal on Image and Video Processing* **2014**(1), 1–13 (2014).

[9] Stone, K. and Keller, J., "Convolutional neural network approach for buried target recognition in fl-lwir imagery," in [*SPIE Defense+ Security*], 907219–907219, International Society for Optics and Photonics (2014).

[10] Lee, E. J., Ko, B. C., and Nam, J.-Y., "Recognizing pedestrians unsafe behaviors in far-infrared imagery at night," *Infrared Physics & Technology* **76**, 261–270 (2016).

[11] Ding, Z., Nasrabadi, N., and Fu, Y., "Deep transfer learning for automatic target classification: Mwir to lwir," in [*SPIE Defense+ Security*], 984408–984408, International Society for Optics and Photonics (2016).

[12] Chevalier, M., Thome, N., Cord, M., Fournier, J., Henaff, G., and Dusch, E., "Low resolution convolutional neural network for automatic target recognition," in [*7th International Symposium on Optronics in Defence and Security*], (2016).

[13] Teutsch, M., Muller, T., Huber, M., and Beyerer, J., "Low resolution person detection with a moving thermal infrared camera by hot spot classification," in [*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*], 209–216 (2014).

[14] Breckon, T. P., Han, J. W., and Richardson, J., "Consistency in multi-modal automated target detection using temporally filtered reporting," in [*SPIE Security+ Defence*], 85420L–85420L, International Society for Optics and Photonics (2012).

[15] Crawford, S., Craig, R., Haining, A., Parsons, J., Costard, E., Bois, P., Gauthier, F.-H., and Cocle, O., "Thales long-wave advanced ir qwip cameras," in [*Defense and Security Symposium*], 62060H–62060H, International Society for Optics and Photonics (2006).

[16] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints* **abs/1605.02688** (May 2016).

[17] Dieleman, S., Schlüter, J., Raffel, C., Olson, E., and Sønderby, S. K., "Lasagne: First release.," (Aug. 2015).

[18] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," in [*Advances in neural information processing systems*], 1097–1105 (2012).

[19] Glorot, X. and Bengio, Y., "Understanding the difficulty of training deep feedforward neural networks.," in [*Aistats*], **9**, 249–256 (2010).

[20] Nair, V. and Hinton, G. E., "Rectified linear units improve restricted boltzmann machines," in [*Proceedings of the 27th International Conference on Machine Learning (ICML-10)*], 807–814 (2010).

[21] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R., "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580* (2012).

[22] Bridle, J. S., "Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition," in [*Neurocomputing*], 227–236, Springer (1990).

[23] Duchi, J., Hazan, E., and Singer, Y., "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research* **12**(Jul), 2121–2159 (2011).